Multidiszciplináris tudományok, 13. kötet. (2023) 1 sz. pp. 141-153 **https://doi.org/10.35925/j.multi.2023.1.11**

ROBUST CLUSTERING BASED ON THE MOST FREQUENT VALUE METHOD

Ferenc Tolner
Pannon Business Network Association
Óbuda University Doctoral School of Applied Informatics and Applied Mathematics
32–36 Zanati út, Szombathely, H-9700, e-mail: ferenc.tolner@am-lab.hu

Sándor Fegyverneki
Department of Applied Mathematics, Institute of Mathematics, University of Miskolc
Miskolc-Egyetemváros, H-3515, e-mail: matfs@uni-miskolc.hu

Balázs Barta
Pannon Business Network Association
32–36 Zanati út, Szombathely, H-9700, e-mail: balazs.barta@pbn.hu

György Eigner
University Research and Innovation Center, Physiological Controls Research Center, Óbuda University
Biomatics and Applied Artificial Intelligence Institution,
John von Neumann Faculty of Informatics, Óbuda University
96/b Bécsi út, Budapest, H-1034, e-mail: eigner.gyorgy@nik.uni-obuda.hu

Abstract

Assigning observations to highly separable although relatively homogeneous groups is still a challenging task despite the abundance of well-elaborated theories and effective, practical algorithms. Not just the aim of clustering then the underlying data itself influences the choice of method and the way of assessing the results. Outliers and non-normal data distribution can lead to surprising, unstable and many times undesirable clustering results especially in higher dimensions. This implies the importance of some human supervision in case of such unsupervised algorithms as well. In this paper a robust clustering alternative is presented based on the Most Frequent Value Method for crisp-type clustering in case of real-life data. The proposed approach is compared with the k-Medians algorithm. A favourable attribute of the applied procedure is its ease of application on multidimensional data sets where critical judgment of formed groups is particularly troublesome.

Keywords: *Most Frequent Value, k-MFVs, outlier map, robust clustering, anomaly detection*

1. Introduction

Data mining corresponds to the information extraction process from data that might be previously unknown and non-trivial. The applied methods are mostly descriptive or predictive in nature. Clustering, the unsupervised ordering of unlabelled data into separable, homogeneous groups belongs to the latter one, which is often used in itself or as an intermediate step of pre-procession (Pardeshi et al., 2010). A major goal of clustering is to split big data sets into smaller, uniform segments based on similarity

features that can be further investigated in smaller scale (Rousseeuw et al., 2011). Thereby the hidden information can be broken down into smaller units that is easier to interpret (Kirshners et al., 2012; Thommassey et al., 2006).

Outliers, noises and non-sharp cluster borders may however pose challenges to many clustering algorithms. Therefore, robustification of algorithms when working with real-life data is of particular importance in order to be able to stabilize efficiency and predictive power (Dorabiala et al., 2021). Nonetheless, robustness and stability depend not just on the underlying data then on the clusters themselves. Outliers often form heterogeneous groups with the bulk of the data, therefore clustering can theoretically isolate them (Syarif et al., 2012; Jiang et al., 2016; Chitrakar et al., 2012). However, it is not necessarily worth to seek outliers this way, since noisy observations can lead to "chaining effects" via "bridging points" that can result the density-based interconnection of different data groups and low breakdown points (The percentage amount of outlier points that leads to an unacceptable classification of the data points.) of clustering algorithms (e.g.: k-means) (García-Escudero et al., 2010).

In the course of clustering, outliers and separate groups can be identified without any prior knowledge about the data. Oftentimes different approaches are combined in literature (e.g.: partition based-, hierarchical-, density based-, grid based or model-based techniques) to reach higher efficiency but there are limitations of validity depending on size- and nature of data, number of dimensions, distributions etc. (Chitrakar et al., 2012). Thereby, the selection of appropriate algorithm shall be done accordingly and by no means automatically (Arbin et al., 2015).

Authors of related works often build upon robust approaches like k-Medoids / k-Medians that enjoys widespread popularity (Syarif et al., 2012; Soni et al., 2017; Shamsuddin et al., 2019; Velmurugan et al., 2015; Madhulatha, 2011; Aryuni et al., 2018; Widiyaningtyas et al., 2019; Dharmarajan et al., 2016; Drias et al., 2016). In general, it shows higher accuracy than the well-known k-Means, but its run-time increases fast with sample size that makes it unfavourable in case of bigger problems (Arbin et al., 2015; Olukanmi et al., 2020). On the other hand robust statistical methods are gaining more attention, which instead of only focusing on robust location parameters try to rely on the "bulk" of the data by performing adequate trimming or suppressing of "far-lying" observations (Dorabiala et al., 2021; García-Escudero et al., 1999, 2003 and 2010). A great advantage of the latter approach is that it enables higher-dimensional investigations as well and can be extended to PCA or multivariate outlier detection (Rousseeuw et al., 1990 and 2018; Hubert et al., 2008). Due to these attractive aspects the present study would like to further contribute to robust statistical investigations of the wide field of cluster analysis with the application of the *Most Frequent Value Method* (MFV) developed by (Steiner et. al. 1997) that enables the usage of robust location- and scale parameters without discarding any – possibly valuable – data.

The rest of the paper is structured as follows: In Sec. II the overall concept of MFV-robustified clustering is outlined, then in Sec. III the investigated data sets and concept of outlier identification is presented. The gained comparative results based on various metrics of evaluation are then synthesized in Sec. IV and finally, an outlook is given for further improvement possibilities with the main conclusions in Sec. V.

2. Theoretical background

In various walks of life (e.g.: biology, economy etc.) outliers cannot be regarded simply as measurement errors, anomalies or as members of different populations, since little is known about the mechanics of the unknown model in the background. In many practical fields data points are often too valuable just

to be removed from the original population. Furthermore, it is typically required to be able to unambiguously attach labels to observations and describe general cluster attributes (*crisp* clustering).

The hereby outlined MFV-robustified clustering approach relies on the well-known Lloyd's algorithm (Arthur et al., 2007) by calculating a robust location parameter (*MFV* or *Most Frequent Value* [Not identical with the mode of a data distribution that corresponds to the location of local maxima and therefore might not be unique.]) as cluster centroid simultaneously with a robust scale parameter (called *dihesion*) of the formed clusters in each iteration step (see Algorithm 1). Unlike trimming procedures all observations are taken into consideration but with different weights corresponding their location within the data distribution of each group.

Algorithm 1: (Pseudocode of k-MFVs clustering)
Data:
k: number of clusters
D: a set of objects of cardinality n
1. Initialization of centroids $(c_1, c_2, \ldots, c_k)$ in D

while no changes in cluster centroids
2.1. for each data point x_i calculate distance from centroid in each cluster
2.2. assign objects to cluster with nearest centroid (swapping step)
2.3. for each cluster $j = 1..k$ recalculate centroids as MFV value of each actual cluster
end
return Centroids, cluster labels, dihesion values

The MFV value (Eq. 1.) and *dihesion (*Interpreted as the reciprocal of data-cohesion, which characterizes the spread of the data.) (Eq. 2.) of a 1D data distribution $x_i, i \in [1, n)$ can be calculated according to the following joint iterative procedure:

$$M_{k,x} = \frac{\sum_{i=1}^{n} \frac{(k\epsilon)^2}{(k\epsilon)^2 + (x_i - M_{k,x})^2} \cdot x_i}{\sum_{i=1}^{n} \frac{(k\epsilon)^2}{(k\epsilon)^2 + (x_i - M_{k,x})^2}} \tag{1}$$

$$\epsilon^2 = \frac{3 \cdot \sum_{i=1}^{n} \frac{(x_i - M_{k,x})^2}{(\epsilon^2 + (x_i - M_{k,x})^2)^2}}{\sum_{i=1}^{n} \frac{1}{(\epsilon^2 + (x_i - M_{k,x})^2)^2}} \tag{2}$$

For the initialization of the iteration usually the median is used as the starting value of the MFV, while the MAD (Median Absolute Deviation.) for the *dihesion*. The above constituting equations can also be derived based on the minimization of the Kullback–Leibler information divergence and provide the $M_{k,x}$ location parameter as a weighted average. The weights calculated from a Cauchy distribution increases outlier resistance by down-weighting the far-lying observations measured from the location of data concentration. In order to enhance and tune statistical efficiency without prior knowledge of the data distribution type at hand a simple constant is introduced by Steiner et. al., which is advised to be

$k = 2$ for a wide range of distributions occurring in practice (based on extensive Monte Carlo simulations) (Steiner et al., 1997). By adjusting the k constant the amount of data taken into consideration with higher weight can be tuned and selected according to the specific data distributions. In case of increasing the tuning parameter more data points will be ordered to the "bulk" of the data and to the close-lying outliers higher weights will be assigned when calculating the MFV and *dihesion* values.

It can be proven that the previous equation system is equivalent with the minimization of Eq. 3. (also called as *P-norm*):

$$G(\epsilon, M_{k,x}) = \sum_{i=1}^{n} ln\left[(M_{k,x} - x_i)^2 + (k\epsilon)^2\right] \tag{3}$$

This provides an opportunity for further statistical procedures also for multivariate cases that are more robust and outlier resistant than those based on the minimization of the L2- or even L1-norm for a wide range of distributions "far from" the Gaussian (Steiner et al., 1991 and 1997).

Robustificating Lloyd's algorithm by using *Most Frequent Values* as cluster centroids is made in order to construct an alternative to the numerous unsupervised methodologies that perform the breaking-down of data into smaller parts. With this approach it can be achieved not to discard any data point prior to the clustering, since every data point may represent valuable information and without a proper model (e.g.: economic processes) on the background processes or knowledge about the various error types, dropping data is less recommended. Therefore, our present focus is on *crisp*-like clustering approaches with no data exclusion.

Since only the selection of centroids is modified, we expect to have somewhat similar clustering results as provided by well-known k-Means and k-Medians algorithms. However, due to the iterative depiction of the MFV values an increased time consumption is expected. Moreover, the MFV-robustified alternative (k-MFVs) will also predict spherical-shaped clusters in the multivariate space that instantaneously offers future development directions towards considering elliptical-shaped clusters.

3. Experimental setup

The alteration of centroid calculation shall lead to different classification of the data and new centroid coordinates as well. Thereby, the accuracy and interpretation of the grouping could be different compared to k-Means and k-Medians. In order to look into this, we investigate the k-Means and k-Medians algorithms alongside with the outlined k-MFVs in case of 4 real-life data sets accessible at the UCI database (Dua et al., 2017). Data sets with known classification and relatively small cluster sizes have been selected with different sample sizes and feature numbers for investigation in other literature sources as well (Pérez-Ortega et al., 2017). The main characteristics of the investigated data sets are listed in Table 1.

The *Long Jump* data set contains the results of two long jump trials from the 1988 Olympic Games of men decathlon and women heptathlon. Being a set of one-dimensional observations, it is adequate for visual comparison of different clustering methods. On Fig. 1 besides the original data the results of k-Means, k-Medians, trimmed k-Means at $\alpha = 0.05/02$ levels and the proposed k-MFVs ($k = 2$) are presented in case of the presence of a single outlier that represents a disqualified jump (therefore with zero value) (García-Escudero et al., 1999).

Table 1. *Dataset information*

Dataset name	Sample size	No. of features	No. of clusters	Distribution per cluster
Long Jump	58	1	2	33-25
Iris	150	4	3	50-50-50
Wine	178	13	3	71-59-48
Ecoli	336	7	4	143-116-52-25
Breast Cancer	569	30	2	357-212

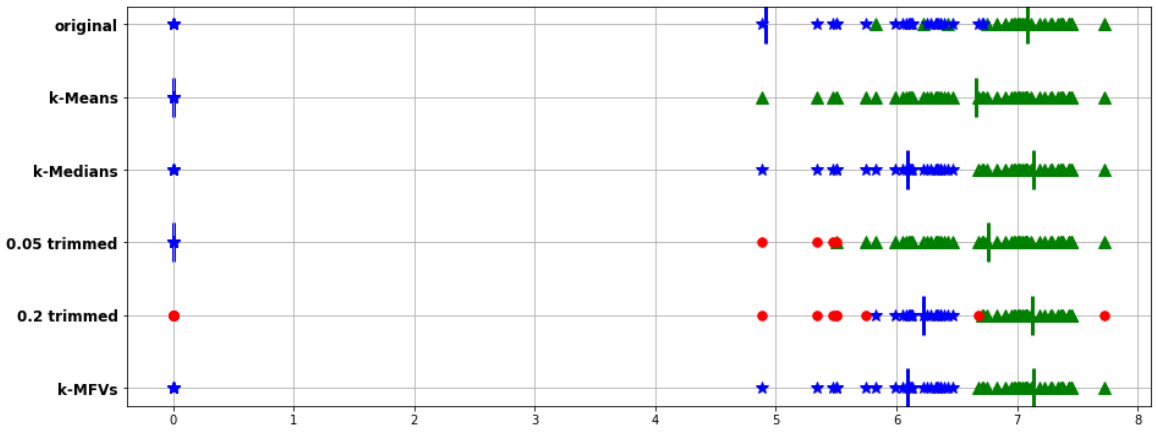

Figure 1. *Comparison of clustering algorithms in case of the "Long Jump" data in the presence of a single outlier*

The original data show some overlap among men's and women's outcomes and the group mean of that containing the single outlier is highly biased towards it. This overlap of the two groups cannot be differentiated by any of the investigated algorithms. Moreover, the presence of the disqualified data resulted the k-Means and the trimmed k-Means ($\alpha = 0.05$) to break down. At the same time, k-Medians and k-MFVs ($k = 2$) and trimmed k-means ($\alpha = 0.2$) proved to be resistant enough and the trimmed k-Means served with additional information on the identified outliers.

Since our aim is to cluster the data in the presence of outliers without discarding them the robust Mahalanobis distances are used to specify outliers in the formed groups. The empirical- and robust within-group Mahalanobis distances can be compared to the critical value of $\sqrt{\chi^2_{1,0975}}$ suggested by (Hubert et al., 2008) in Fig. 2 shows the calculated distances for the two resulted groups in case of k-MFV ($k = 2$) clustering. The points above the critical values can be considered as group-wise outliers and their further investigation can be done subsequently. A main advantage is that no data had to be suspended, thus the centroids did not get biased because of that. The applied methodology is easy to interpret and can further be extended for higher dimensional investigations. This possibility holds for the 4 UCI datasets; however the detailed investigation of outliers is beyond the scope of the present study.

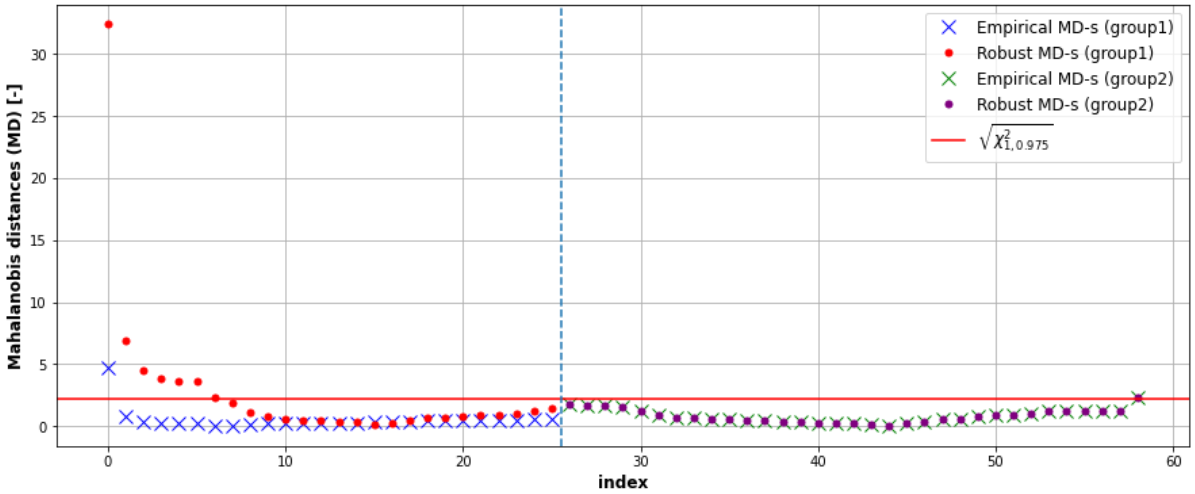

Figure 2. *Empirical- and robustified Mahalanobis distances for the "Long Jump" data extended with a single outlier in case of the k-MFVs (k = 2) partitioning into two separate groups as a function of ordered element indices*

Throughout the experimental investigation all of the multidimensional data were used in their "raw" form without dimensional reduction or standardisation. Since initialization is crucial, k-Means++ and DBSCAN were implemented and tested in order to avoid randomness in the resulted clusters. The k-Means++ proved to be improper for our purposes, because the aim of the robustification with the *Most Frequent Value* is to cluster the "bulk" of the data and place centroids around high density locations, but k-Means++ initialization typically led the algorithms to stuck in isolated outlying groups. Therefore, DBSCAN has been selected and by proper parameter sweeping the desired number of initial centroids were defined for each dataset in a reproducible way.

4. Experimental results and validation

The gained results of the "k-MFVs" algorithm in case of the five selected data sets are outlined in Table 2. together with the k-Means' and k-Medians' for comparative purposes. As important metrics for the judgement of the algorithms the number of "swap-s" (number of iteration until no more changes in centroids – and point assignments – is achieved) and computational time for convergence have been recorded besides five clustering validity indices. The table contains the results only for the optimal cluster numbers known in advance from the labelled datasets (see Table 1).

From Table 2 it can be seen that the computational time for the k-MFVs algorithm is much higher, however it cannot be directly compared, since it highly depends on the implementation of the applied built-in functions. Therefore, the relative increase of the time required has been inspected as a function of dataset size and cluster numbers. According to the results it is not straightforward to expect a time increment with increasing sample size, rather the number of elements in each cluster plays an essential role. For the *Ecoli*, which was the second largest investigated dataset we gained an order of magnitude smaller computational times as in case of the *Breast Cancer*, while the k-Means and k-Medians performed in the same order of magnitude, albeit these required somewhat better run times as well. The latter data had only two relatively large clusters, while the former four clusters out of which three was

relatively small in cardinality. Therefore, k-MFVs is expected to serve more cost-efficient result in case of large data sets with more clusters.

The calculation of the MFV values according to Eq. 1 and Eq. 2 are rather time consuming. Throughout our investigations the implementation was done in *Python 3.7.13* within an *Anaconda* framework (Anaconda 2-2.4.0., 2016), where the exit criterion from the iterative procedure was established in $\Delta\epsilon_{max} < 10^{-5}$ for the dihesion values in two subsequent iterations. This is relatively strict and its necessity might be data dependent, therefore could be loosen up in order to gain significant time reduction for the k-MFVs algorithm at the same clustering accuracy. For different convergence trajectories of the MFV iterative algorithm in case of the Long *Jump data* set see Fig. 3.

The number of centroid swap-s and data reassignments to the groups (N_{swap}) did not showed significant variance in case of the algorithms. In case of the *Wine* dataset the k-Medians required more swaps than the others, however in general the k-MFVs resulted in swap numbers between the swap numbers of the k-Means and k-Medians. The k-Medians and k-MFVs needed approximately the same number of swaps for higher cluster numbers ($n = 5, 6, 7$), nonetheless k-Medians performed outstandingly in this aspect for the *Ecoli* at these cluster number choices. For instance, for $n = 7$ (that is far from the optimal clustering setting) the k-Medians required only 5 swap-s, while the k-MFVs 28 and the k-Means 30.

In higher dimensions data are hard to visualise and cluster validity indexes can be used to rely on in order to control the resistance and robustness of the applied procedure in case of specific data. By the investigation of these indexes different methods can be compared and/or optimal cluster numbers can be sought.

Nevertheless, literature draws attention to the possible dependence of such metrics on the selected clustering algorithms, since noises and outliers might influence them even in cases when their presence does not result significantly different groups (Wu et al., 2009). As a non-sensitive validity index to clustering algorithms the S_{dbw} metric has been selected that has to be minimized in order to gain an optimal grouping (Liu et al., 2010; Halkidi et al., 2001).

Whereas the labelling information was also given for all the datasets, Silhouette (SC)-, Davis-Bouldin-indices (DBI) were also calculated besides Adjusted Mutual Information (AMI) (Drias et al., 2016; Pérez-Ortega et al., 2017) and Rand indices (R) (Olukanmi et al., 2020; Pérez-Ortega et al., 2017; Pratap et al., 2011; Cerioli et al., 2018) to compare the resulted groups with the known labels. The SC- (Shamsuddin et al., 2019), AMI- and R indices had to be maximized while the S_{dbw}- and DBI (Aryuni et al., 2018) metrics had to be minimized in the function of cluster numbers (see Table 2). Nevertheless, k-Medians and k-MFVs performed slightly better in all of the investigated cases at most of the parameter settings, k-Means was able to serve with better results in case of *Ecoli* and *Breast Cancer* datasets in terms of DBI or SC, however the differences could only be measured in the third digit.

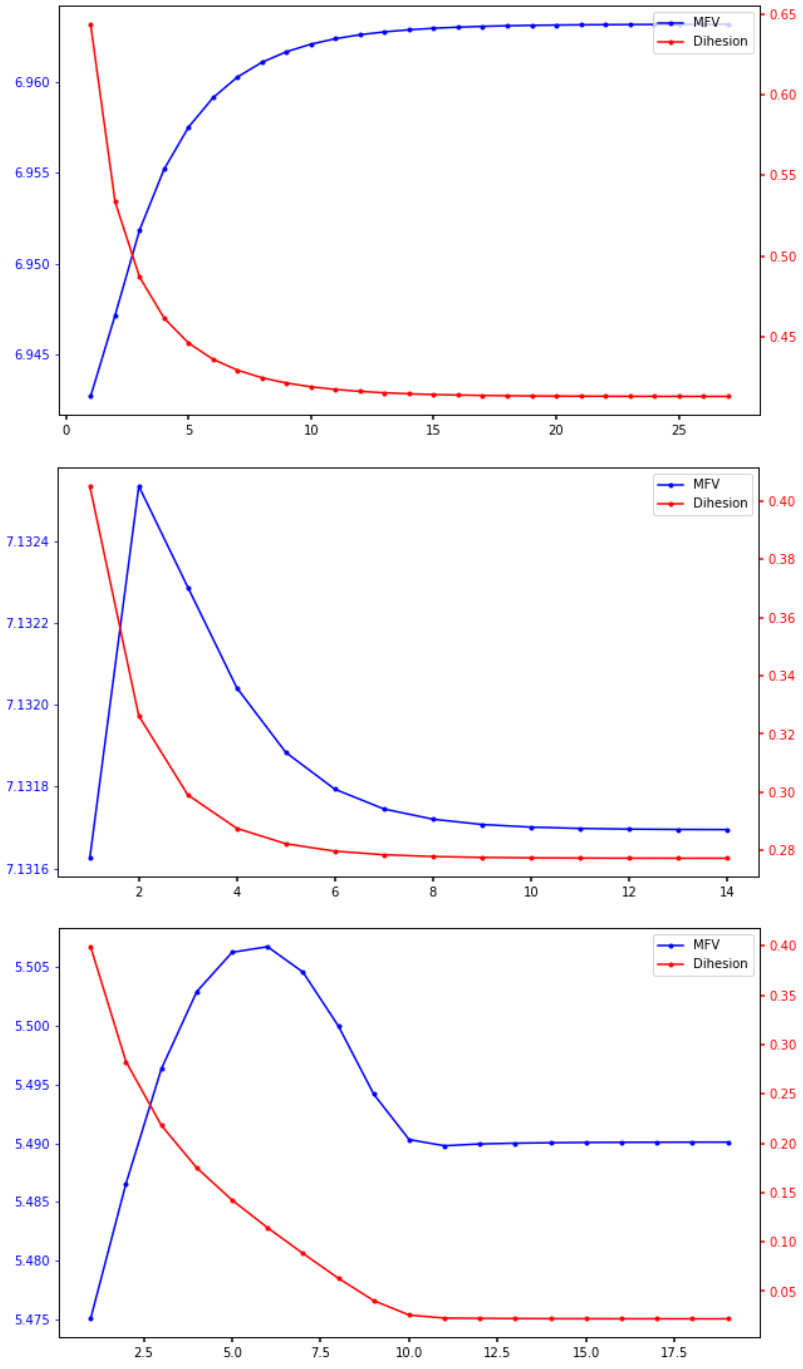

Figure 3. *Typical trajectories of the "Most Frequent Value" and "dihesion" towards convergent state in arbitrary cases.*

Table 2. *Validity indices, main performance metrics and resulted cluster distributions for different location parameter choices and tuning parameter setting (noteworthy results per row indicated by bold)*

	Mean	**Median**	**MFV (k = 1)**	**MFV (k = 2)**	**MFV (k = 3)**	**MFV (k = 4)**
	"Long Jump" dataset ($n_{clust} = 2$)					
N_{swap}	4	4	4	4	4	4
t(s)	0.0010	0.0020	0.0479	0.0568	0.0479	0.0409
R	0.6225	0.6225	0.6225	0.6225	0.6225	0.6225
S_{dbw}	0.6511	0.6511	0.6511	0.6511	0.6511	0.6511
AMI	0.5087	0.5087	0.5087	0.5087	0.5087	0.5087
SC	0.6645	0.6645	0.6645	0.6644	0.6644	0.6644
DBI	0.4754	0.4754	0.4754	0.4754	0.4754	0.4754
Clusters	33-25	33-25	33-25	33-25	33-25	33-25
	"Iris" dataset ($n_{clust} = 3$)					
N_{swap}	6	6	6	6	6	6
t(s)	0.0030	0.0040	0.6207	0.5513	0.6477	0.5505
R	0.7302	0.7439	**0.7437**	0.7304	0.7302	0.7302
S_{dbw}	0.3356	0.3373	0.3373	0.3356	0.3356	0.3356
AMI	0.7551	0.7631	**0.7631**	0.7551	0.7551	0.7551
SC	0.5526	0.5509	0.5509	0.5526	**0.5551**	0.5526
DBI	0.6623	0.6662	**0.6662**	0.6623	0.6623	0.6623
Clusters	50-62-38	50-61-39	50-61-39	50-62-38	50-62-38	50-62-38
	"Wine" dataset ($n_{clust} = 3$)					
N_{swap}	13	19	12	11	11	12
t(s)	0.0070	0.0119	6.0123	4.4153	4.0440	4.5023
R	0.3518	**0.3715**	0.3389	0.3415	0.3415	0.3415
S_{dbw}	0.4092	**0.3746**	0.4134	0.4139	0.4139	0.4139
AMI	**0.4168**	0.4131	0.4068	0.4093	0.4093	0.4093
SC	0.5596	**0.5708**	0.5479	0.5447	0.5447	0.5447
DBI	0.5496	**0.5317**	0.5531	0.5541	0.5447	0.5541
Clusters	49-27-102	62-48-68	48-30-100	47-31-100	47-31-100	47-31-100
	"Ecoli" dataset ($n_{clust} = 4$)					
N_{swap}	7	7	7	8	**6**	**6**
t(s)	0.0049	0.0070	1.7726	1.9345	1.4765	**1.4944**
R	0.6847	0.6861	0.7541	0.6764	**0.7619**	**0.7619**
S_{dbw}	0.6607	0.6607	0.6608	0.6608	0.6607	0.6606
AMI	0.6416	0.6483	0.6765	0.6353	**0.6836**	**0.6836**
SC	0.4221	0.4210	0.4210	0.4210	0.4206	**0.4226**
DBI	**0.9403**	0.9428	0.9423	0.9423	0.9403	0.9403
Clusters	149-104-75-8	149-103-76-8	148-104-76-8	148-104-76-8	149-104-75-8	149-104-75-8

	Mean	**Median**	**MFV (k = 1)**	**MFV (k = 2)**	**MFV (k = 3)**	**MFV (k = 4)**
	"Breast Cancer" dataset ($n_{clust} = 2$)					
N_{swap}	10	**6**	**6**	8	8	7
t(s)	0.0076	0.0050	16.6334	20.1386	20.2572	17.2017
R	0.4914	**0.5338**	0.5286	**0.5338**	0.5124	0.5019
S_{dbw}	0.7912	0.7857	**0.7854**	0.7857	0.7881	0.7895
AMI	0.4640	**0.4973**	0.4839	**0.4973**	0.4805	0.4722
SC	**0.6973**	0.6921	0.6911	0.6921	0.6952	0.6965
DBI	**0.5044**	0.5139	0.5154	0.5139	0.5087	0.5064
Clusters	438-131	430-139	429-140	430-139	434-135	436-133

The calculated validity indices showed a rather uniform layout for the different cases. This might indicate that the chosen data are not perfectly suitable for spherical partitioning approaches. The emerged cluster sample sizes further support this statement. For the previously known optimal cluster numbers the sample size distribution resulted to be approximately the same for the *Long Jump*, *Iris* and *Ecoli* datasets. In case of *Wine* data k-Medians led to a more similar cluster distribution to the known one in alignment with the better validity indices. For the *Breast Cancer* data k-Medians and k-MFVs with $k = 2$ selection provided the same accuracy.

5. Conclusions, future work

In the present study a clustering algorithm based on Lloyd's algorithm has been investigated in case of real-life data. As cluster centroids the *Most Frequent Value* was selected that is a robust and outlier resistant location parameter of a data distribution. Albeit the current results showed a significant increase in run-time requirement compared to k-Means and k-Medians the gained accuracy measured by various metrics can be considered as encouraging.

The motivation for the robustified crisp-type algorithm creation was to further enrich the selection of robust clustering methods with an alternative that do not expel any data point by judging it an outlier. This is of paramount importance with regard of our future research where economic data is to be investigated by breaking it down into smaller chunks via clustering. In such cases every data point represents valuable information and by neglecting them the variability of the data would be distorted and the derived results biased.

As a future work we would like to decrease the runtime of the algorithm and perform further comparative studies on artificial- and real-life data where multidimensional data distributions with high skewness are present and – according to our expectations – the k-MFVs might score better. Similarly to the outlined comparative study with the k-Medians algorithm we would like to use the resulted dihesion values for calculating robust group-wise Mahalanobis distances and consider ellipsoid-shaped clusters.

6. Acknowledgements

The publication of this article has been supported by the Robotics Special College via the "NTP-SZKOLL-21-0034 Talent management and the professional community building at the ÓE ROSZ" project, the Applied Informatics and Applied Mathematics Doctoral School of Óbuda University, and the the Pannon Business Network Association.

References

[1] Arbin, N., Suhaimi, N. S., Mokhtar, N. Z. and Othman, Z. (2015). Comparative analysis between K-Means and K-Medoids for statistical clustering. *3rd International Conference on Artificial Intelligence, Modelling and Simulation (AIMS)*, pp. 117–121. **https://doi.org/10.1109/AIMS.2015.82**

[2] Arthur, D. and Vassilvitskii, S. (2017). k-means++: The advantages of careful seeding. *Proceedings of the 18th Annual ACM-SIAM Symposium on Discrete Algorithms*.

[3] Aryuni, M., Madyatmadja, E. D. and Miranda, E. (2018). Customer segmentation in XYZ Bank using K-Means and K-Medoids clustering. *International Conference on Information Management and Technology (ICIMTech)*, **https://doi.org/10.1109/ICIMTech.2018.8528086**

[4] Cerioli, A., Farcomeni, A. and Riani, M. (2018). Wild adaptive trimming for robust estimation and cluster analysis. *Scandinavian Journal of Statistics*, 46(1), pp. 235–256. **https://doi.org/10.1111/sjos.12349**

[5] Chitrakar, R. and Chuanhe, H. (2012). Anomaly detection using support vector machine classification with k-Medoids clustering. *Third Asian Himalayas International Conference on Internet*, **https://doi.org/10.1109/AHICI.2012.6408446**

[6] Computer software. Vers. 2-2.4.0. Anaconda, *Anaconda Software Distribution*, Web.: **https://anaconda.com**, Nov. 2016.

[7] Dharmarajan, A. and Velmurugan, T. (2016). Efficiency of k-Means and k-Medoids clustering algorithms using Lung Cancer Dataset. *International Journal of Data Mining Techniques and Applications*, 5(2), pp. 150–156, **https://doi.org/10.20894/IJDMTA.102.005.002.011**

[8] Dorabiala, O., Kutz, J. N. and Aravkin, A. Y. (2021). *Robust Trimmed kmeans*, ArXiv, vol. abs/2108.07186.

[9] Drias, H., Cherif, N. F. and Kechid, A. (2016). *k-MM:* A hybrid clustering algorithm based on k-Means and k-Medoids. *Advances in Nature and Biologically Inspired Computing*, pp. 37–48. **https://doi.org/10.1007/978-3-319-27400-3_4**

[10] Dua, D. and Graff, C. (2017). *UCI machine learning repository*. [Online]. Available: **http://archive.ics.uci.edu/ml**

[11] García-Escudero, L. A. and Gordaliza, A. (1999). Robustness properties of k Means and trimmed k Means. *Journal of the American Statistical Association*, 94(447), pp. 956–969. **https://doi.org/10.2307/2670010**

[12] García-Escudero, L. A., Gordaliza, A. and Matrán, C. (2003). Trimming tools in exploratory data analysis. *Journal of Computational and Graphical Statistics*, 12(2), pp. 434–449. **https://doi.org/10.1198/1061860031806**

[13] García-Escudero, L. A., Gordaliza, A., Matrán, C. and Mayo-Iscar, A. (2010). A review of robust clustering methods. *Advances in Data Analysis and Classification*, 4(2), pp. 89–109. **https://doi.org/10.1007/s11634-010-0064-5**

[14] Halkidi, M. and Vazirgiannis, M. (2001). *Clustering validity assessment: finding the optimal partitioning of a data set*. Proceedings 2001 IEEE International Conference on Data Mining. **https://doi.org/10.1109/ICDM.2001.989517**

[15] Hubert, M., Rousseeuw, P. J. and Aelst, S. V. (2008). High-breakdown robust multivariate methods. *Statistical Science*, 23(1), pp. 92–119, **https://doi.org/10.1214/088342307000000087**

[16] Jiang, F., Liu, G., Du, J. and Sui, Y. (2016). Initialization of K-Modes clustering using outlier detection techniques. *Information Sciences*, 332(1), pp. 167–183. **https://doi.org/10.1016/j.ins.2015.11.005**

[17] Kirshners, A., Borisov, A. and Parshutin, S. (2012). Robust cluster analysis in forecasting task. *International Conference on Applied Information and Communication Technologies*.

[18] Liu, Y., Li, Z., Xiong, H., Gao, X. and Wu, J. (2010). Understanding of internal clustering validation measures. *IEEE International Conference on Data Mining*. **https://doi.org/10.1109/ICDM.2010.35**

[19] Madhulatha, T. S. (2011). Comparision between k-Means and k-Medoids clustering algorithms. *International Conference on Advances in Computing and Information Technology Advances in Computing and Information Technology (ACITY 2011), Communications in Computer and Information Science*, vol. 198, pp. 472–481, **https://doi.org/10.1007/978-3-642-22555-0_48**

[20] Olukanmi, P., Nelwamondo, F. and Marwala, T. (2020). Effect of data parameters and seeding on k-Means and k-Medoids. *International Conference on Artificial Intelligence, Big Data, Computing and Data Communication Systems (icABCD)*. **https://doi.org/10.1109/icABCD49160.2020.9183892**

[21] Pardeshi, B. and Toshniwal, D. (2010). Improved K-Medoids clustering based on Cluster Validity Index and object density. *IEEE 2nd International Advance Computing Conference (IACC)*, **https://doi.org/10.1109/IADCC.2010.5422924**

[22] Pratap, A. R., Siddhartha, P. V. P., Devi, J. R. and Vani K. S. (2011). An efficient density based improved K- Medoids clustering algorithm. *International Journal of Advanced Computer Science and Applications*, 2(6), **https://doi.org/10.14569/IJACSA.2011.020607**

[23] Pérez-Ortega, J., Almanza-Ortega, N. N., Adams-López, J., González-Gárcia, M., Mexicano, A., Saenz-Sánchez, S. and Rodríguez-Lelis, J. (2017). Improving the efficiency of the K-medoids clustering algorithm by getting initial medoids. *Advances in Intelligent Systems and Computing*, vol. 569, **https://doi.org/10.1007/978-3-319-56535-4_13**

[24] Rousseeuw, P. J. and Hubert, M. (2018). Anomaly detection by robust statistics. *WIREs Data Mining and Knowledge Discovery*, 8(2), e1236, **https://doi.org/10.1002/widm.1236**

[25] Rousseeuw, P. J. and Hubert, M. (2011). *Robust statistics for outlier detection*. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, vol. 1. **https://doi.org/10.1002/widm.2**

[26] Rousseeuw, P. J. and van Zomeren, B. C. (1990). Unmasking multivariate outliers and leverage points. *Journal of the American Statistical Association*, 85(411), pp. 633–639. **https://doi.org/10.1080/01621459.1990.10474920**

[27] Shamsuddin, N. R. and Mahat, N. I. (2019). Comparison between k-Means and k-Medoids for mixed variables clustering. *Proceedings of the Third International Conference on Computing, Mathematics and Statistics (iCMS2017)*, pp. 303–308. **https://doi.org/10.1007/978-981-13-7279-7_37**

[28] Soni, K. G. and Patel, A. (2017). Comparative analysis of K-means and Kmedoids algorithm on IRIS data. *International Journal of Computational Intelligence Research*, 13(5), pp. 899–906.

[29] Steiner, F. (Ed.) (1997). *Optimum Methods in Statistics*. Akadémiai Kiadó, Budapest, Hungary, 370 p., ISBN: 963 05 7439 X.

[30] Steiner, F. (Ed.) (1991). *The most frequent value. Introduction to a modern conception of statistics*. Akadémiai Kiadó, Budapest, Hungary, 315 p., ISBN: 963 05 5687 1.

[31] Syarif, I., Prugel-Bennett, A. and Wills, G. (2012). Unsupervised clustering approach for network anomaly detection. *Fourth International Conference on Networked Digital Technologies*. **https://doi.org/10.1007/978-3-642-30507-8_13**

[32] Thomassey, S. and Fiordaliso, A. (2006). A hybrid sales forecasting system based on clustering and decision trees. *Decision Support Systems*, 42(1), pp. 408–421. **https://doi.org/10.1016/j.dss.2005.01.008**

[33] Velmurugan, T. and Dharmarajan, A. (2015). Clustering Lung Cancer data by k-Means and k-Medoids algorithms. *International Conference on Information and Convergence Technology for Smart Society.*

[34] Widiyaningtyas, T., Pujianto, U. and Prabowo, M. I. W. (2019). K-Medoids and K-Means clustering in high school teacher distribution. *International Conference on Electrical, Electronics and Information Engineering (ICEEIE),* **https://doi.org/10.1109/ICEEIE47180.2019.8981466**

[35] Wu, K.-L., Miin-ShenYang, and Hsieh, J.-N. (2009). Robust cluster validity indexes. *Pattern Recognition*, 42(11), pp. 2541–2550, **https://doi.org/10.1016/j.patcog.2009.02.010**